J Risk Uncertain (2018) 56:1–17
https://doi.org/10.1007/s11166-018-9274-6

Complexity in risk elicitation may affect the conclusions: A demonstration using gender differences

Gary Charness[1] · Catherine Eckel[2] · Uri Gneezy[3,4] · Agne Kajackaite[5]

Published online: 27 March 2018

Abstract The Holt and Laury (*American Economic Review, 92*(5), 1644–1655, 2002) mechanism (HL) is the most widely-used method for eliciting risk preferences in economics. Participants typically make ten decisions with different variance options, with one of these choices randomly chosen for actual payoff. For this mechanism to provide an accurate measure of risk aversion, participants need to understand the choices and give consistent responses. Unfortunately, inconsistent and even dominated choices are often made. Can these mistakes lead to a misrepresentation of economic phenomena? We use gender differences in risk taking to test this question. In contrast to many findings in the literature, HL results typically do not find significant gender differences. We compare the HL approach, where we replicate the lack of significant gender differences, with a simpler presentation of the same choices in which

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11166-018-9274-6) contains supplementary material, which is available to authorized users.

✉ Gary Charness
charness@econ.ucsb.edu

Catherine Eckel
ceckel@tamu.ed

Uri Gneezy
ugneezy@ucsd.edu

Agne Kajackaite
agne.kajackaite@wzb.eu

[1] University of California Santa Barbara, 2127 North Hall, Santa Barbara, CA 93106, USA

[2] Texas A&M University, TAMU 4228, College Station, TX 77845-4228, USA

[3] University of California San Diego, 9500 Gilman Drive, La Jolla, San Diego, CA 92093, USA

[4] CREED, University of Amsterdam, Amsterdam, Netherlands

[5] WZB Berlin Social Science Center, Reichpietschufer 50, D-10785 Berlin, Germany

participants make only one of the ten HL decisions; this simpler presentation yields strong gender differences indicating that women are more risk averse than men. We also find gender differences in the consistency of decisions. We believe that the results found in the simpler case are more reflective of underlying preferences, since the task is considerably easier to understand. Our results suggest that the complexity and structure of the risk elicitation mechanism can affect measured risk preferences. The issue of complexity and comprehension is also likely to be present with elicitation mechanisms in other realms of economic preferences.

Keywords Gender · Risk preferences · Elicitation mechanisms · Complexity · Experiment

JEL Classifications B49 · C91 · C99 · D03 · J16

> "Truth is ever to be found in simplicity, and not in the multiplicity and confusion of things."
>
> — Isaac Newton

1 Introduction

Economic models of risk taking are very specific regarding the general properties of risk preferences and the rational behavior associated with those preferences, but they leave the actual level of risk taking as a free parameter. In order to use these models in practice, it is often necessary to measure the actual level of risk taking of the relevant population. A large empirical literature attempts to measure risk attitudes, with a particular emphasis on incentivized measures implemented in lab and field settings (see Charness et al. 2013 and Holt and Laury 2014 for recent reviews).

One consideration when measuring the level of risk taking of experimental participants is that there could be a relation between the complexity of the measure and the difficulty subjects face in completing the measure. The more detailed the set of decisions that subjects are asked to make, the more informative the measure could be, at least in principle. However, these more detailed decision sets may also be more complex and might result in a higher fraction of participants who do not fully understand the experiment, so that their answers reflect some degree of confusion.[1] Hence, there might be a trade-off between the level of detail of the measure and the noise associated with it. Another problem may arise if the measure introduces noise that actually distorts the elicited parameter. For example, certainty-equivalent elicitation methods such as the Becker-DeGroot-Marschak procedure may bias valuations (Hey et al. 2009; Cason and Plott 2014).

To investigate the effect of the complexity of instructions, we employ the most widely-used risk preference elicitation method, popularized by Holt and Laury (2002, hereafter HL).[2] HL has become the gold standard in economics for risk

[1] See for example, Charness et al. (2013) and Dave et al. (2010).

[2] The original idea was introduced by Miller et al. (1969).

preference elicitation. This approach employs an array of 10 decision tasks presented in rows in the form of a table, and featuring a higher-variance option ("risky") and a lower-variance option ("safe"). The higher-variance option becomes more and more attractive in terms of expected payoff as one proceeds down the rows, until choosing this option in the bottom row is the dominant strategy. Participants make a choice in each of the 10 rows; one choice is selected at random for actual payoff. This task is complex because it includes a number of decisions, structured in a specific way, and because it involves a compound lottery (the random selection of decision, and playing out the selected lottery).

Figure 1 (Table 1 from HL) shows the canonical HL decision frame.[3] The likelihood of the high (low) payoff increases (decreases) as one moves down the table, so that Option B is increasingly attractive with lower and lower rows; a risk neutral person should choose Option A in the top four rows and Option B in the bottom six rows. The row in which one switches from choosing Option A to choosing Option B depends on one's risk preferences.

A minimal required behavior for a participant's set of choices to be consistent with any economic theory is that these choices do not violate first-order stochastic dominance. Gamble D has first-order stochastic dominance over gamble E if for any outcome *x*, D gives at least as high a probability of receiving at least *x* as does E, and for some *x*, D gives a higher probability of receiving at least *x*. In the context of HL, this means that one can cross over to choosing Option B at most once (note that Option A is dominated by Option B in the bottom row).[4] However, many experimental participants make choices that cross multiple times between Option A and Option B and/or choose Option A in the bottom row, indicating a fundamental lack of comprehension—in others words, noise. This multiple-crossing rate differs across experimental populations, ranging from 10% with university students to well over 50% with villagers in developing countries (see Charness et al. 2013 and Charness and Viceisza 2016 for lists). In addition, Filippin and Crosetto (2016) find (in their meta-analysis of HL studies) that women are more likely to exhibit inconsistent choices (17.3% of women and 14.1% of men).[5]

We compare the HL method with a simplified method in which, instead of answering the 10-decision HL task and then participating in a lottery that picks one choice for payment, our participants were faced with only one comparison and made only one choice. That is, a participant in the lab was shown only one of the lines in Figure 1, and was asked to make a choice between Option A and Option B for that line. The instructions for this method are shorter and simpler; we refer to this reduced HL method as the "single-choice" method in this paper.

To test whether the distributions of individual risk preferences elicited using the HL and the single-choice method are qualitatively different, we consider the

[3] The column showing the expected payoff difference was not presented to the laboratory participants, but we present it here for expositional purposes.

[4] It is conceivable that one could choose Option B in every row and thus not violate first-order stochastic dominance, but such a pattern would imply a rather implausibly strong preference for taking financial risks.

[5] Their analysis (of 54 HL studies with 7000 subjects) breaks down inconsistent decisions as follows. For multiple switching only, 8.8% of men and 11.8 of women; for dominated choice at decision 10 only, 1.8% of men and 1.2% of women; and both types 3.6% of men and 4.3% of women.

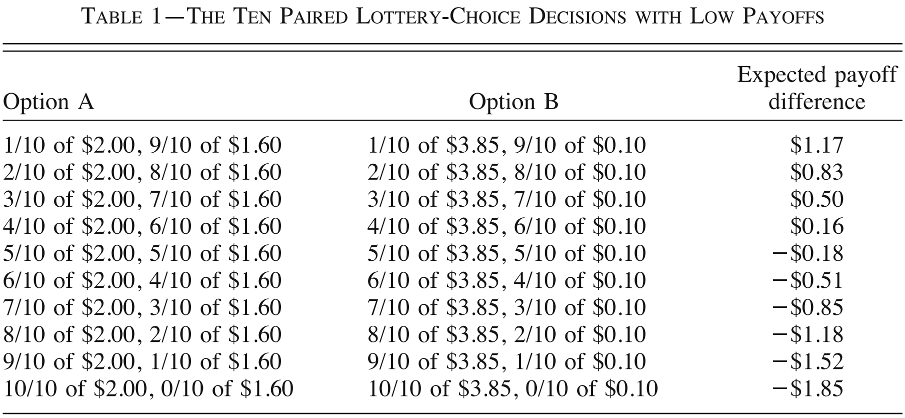

Table 1—The Ten Paired Lottery-Choice Decisions with Low Payoffs

Option A	Option B	Expected payoff difference
1/10 of \$2.00, 9/10 of \$1.60	1/10 of \$3.85, 9/10 of \$0.10	\$1.17
2/10 of \$2.00, 8/10 of \$1.60	2/10 of \$3.85, 8/10 of \$0.10	\$0.83
3/10 of \$2.00, 7/10 of \$1.60	3/10 of \$3.85, 7/10 of \$0.10	\$0.50
4/10 of \$2.00, 6/10 of \$1.60	4/10 of \$3.85, 6/10 of \$0.10	\$0.16
5/10 of \$2.00, 5/10 of \$1.60	5/10 of \$3.85, 5/10 of \$0.10	−\$0.18
6/10 of \$2.00, 4/10 of \$1.60	6/10 of \$3.85, 4/10 of \$0.10	−\$0.51
7/10 of \$2.00, 3/10 of \$1.60	7/10 of \$3.85, 3/10 of \$0.10	−\$0.85
8/10 of \$2.00, 2/10 of \$1.60	8/10 of \$3.85, 2/10 of \$0.10	−\$1.18
9/10 of \$2.00, 1/10 of \$1.60	9/10 of \$3.85, 1/10 of \$0.10	−\$1.52
10/10 of \$2.00, 0/10 of \$1.60	10/10 of \$3.85, 0/10 of \$0.10	−\$1.85

Fig. 1 Holt and Laury (2002) mechanism

gender differences in risk taking found using the two methods. Evidence in the literature shows that, in most risk taking domains, females are more risk averse than males (Eckel and Grossman 2008b; Croson and Gneezy 2009). Charness and Gneezy (2012) collected the gender results from all the experiments that used forms of the Gneezy and Potters (1997) risk elicitation method and that recorded gender. These experiments were not designed to measure gender differences, but rather to answer other economic questions, and used very different instructions, payoffs, and risk levels. The results indicated very robust gender differences in risk taking across the studies, with women almost always being more financially risk averse than men.

Filippin and Crosetto (2016) performed an exercise similar to that in Charness and Gneezy (2012), but with HL data. They find that "…the magnitude of gender differences … is economically unimportant." They further report: "Differences amount to one sixth of a standard deviation, less than a third of the effect found by other elicitation methods (e.g., by Charness and Gneezy, 2012; Eckel and Grossman, 2008a)."

This striking difference in conclusions on gender effects leads one to wonder what could be driving the disparity between the two sets of rather well-established results. A possible explanation is that the complexity of the HL measure causes more participants to be confused and thus leads to noisier results. Clearly, as the measure reflects a significant fraction of inconsistent and dominated choices, any existing underlying gender differences are harder to detect. The structure of the multiple price list might also introduce a bias in elicited preferences, for example, because participants who are slightly confused by the mechanism might have a tendency to switch in the middle of the list. This would tend to compress preferences toward the center-switching point, further masking any gender differences.

In line with Filippin and Crosetto (2016), we find no evidence of a significant gender difference in risk preferences in the standard HL method in either location. However, we see a strong gender difference in risk preference with the simpler single-choice method: females are more risk averse than males. This illustrates how the elicitation method can influence the economic conclusions from an experiment.

One element of our study is that we conduct tests of the HL mechanism with two different sets of instructions, which vary across the level of detail provided and by whether the instructions are read aloud. There is considerably more noisy and inconsistent behavior when the 10-row instructions are simply given in the text and are not read aloud than when examples are provided and the instructions are read aloud. This difference provides a further illustration of the importance of taking complexity into account. When the procedure is complex, it becomes more important to take care to provide complete instructions and to ensure that subjects absorb these instructions.

While we cannot assess inconsistency across choices in our single-choice treatments, the instructions are simpler and easier to understand, and we therefore contend that they are more likely to reflect true underlying preferences (see also Dave et al. 2010; Healy and Brown 2016). We also point out that multiple price lists involve compound lotteries and that compound lotteries are considered by some researchers (e.g., Halevy 2007) to represent ambiguity rather than risk, potentially clouding their use. Further, the structure of the task—observing decisions in a structured list rather than separately—has also been shown to affect elicited preferences (e.g., Cox, Sadiraj, and Schmidt 2015; Healy and Brown 2016). We see this as a cautionary note for employing mechanisms that have a series of within-subject choices and compound lotteries.

The reminder of the paper is organized as follows. Section 2 provides the experimental design and implementation and we present the experimental results in Section 3. Section 4 offers some discussion and concludes.

2 Experimental design and implementation

The experimental design has two treatments. The first is the standard HL method, where participants face a table with 10 decision tasks presented in rows (similar to Figure 1; see the Electronic Supplementary Material). The participant has to choose an Option A or Option B in each of the ten rows, where A is the "safe" option and B is the "risky" option, with higher variance in payoffs. Option A has a higher expected payoff in the first rows; however, as a participant descends down the rows Option B becomes more attractive in expected payoffs. Depending on the risk preferences, a participant should switch from Option A to Option B at the latest in the tenth row, where B is the dominant strategy. While participants make a choice in each of the ten rows, only one row is paid out—a throw of a ten-sided die determines which row is relevant for the actual payoff. A further die roll determines the earnings from the selected row.

In the second approach, instead of showing the participants the full set of HL choices, we show each person only one row of the HL table and ask them to make one binary decision. We selected a subset of the rows (using rows 1, 3, 5, 7, or 9) and vary the row that is shown (see the instructions in the Electronic Supplementary Material). For example, in the first-row treatment, the participant has to choose between Option A, which leads to a payoff of $2 with 10% probability and a payoff of $1.60 with 90% probability, and Option B, where the participant receives $3.85 with 10% chance and $0.10 with 90% chance. That was the only decision the participant had to make in this treatment, which makes the second treatment cognitively less demanding than the multiple price list we use in the first treatment.

Our experimental design involves a further treatment. We conducted the experiment in two laboratories—at the Rady Behavioral Lab at the University of California, San Diego and at the Economics Research Lab at Texas A&M University. In addition, while in the HL treatment in San Diego we used the original HL instructions that were published together with Holt and Laury (2002), in the Texas A&M experiment we used more extensive HL instructions, which explain some rows and also the payoffs' determination in a more detailed way (see the Electronic Supplementary Material). These instructions were adapted slightly from the current version of the Holt and Laury instructions in Veconlab, and are updated (by Holt) from those in the original 2002 paper.[6] Also, at Texas A&M, following standard practice for that lab, we read the HL instructions out loud in order to make it more likely that subjects read and understood the instructions. Therefore, in the HL treatment, we expected to observe fewer inconsistent choices in Texas than in San Diego.[7]

It is worth noting that both sets of instructions emphasize the difference in the tenth choice, but the Texas instructions are more detailed. For example, in San Diego, the HL instructions state, "In fact, for Decision 10 in the bottom row, the die will not be needed since each option pays the highest payoff for sure, so your choice here is between 200 pennies or 385 pennies." The instructions used in Texas similarly note, "For decision 10 shown below, the random die throw will not be needed, since the choice is between amounts of money that are fixed: $2.00 for Option A and $3.85 for Option B." In addition below this sentence there is a table illustrating the two options and making it more obvious that Option A is dominated.

We conducted the experiments in February–July 2015 at the Rady Behavioral Lab, UC San Diego and the Economics Research Laboratory (ERL) at Texas A&M University. Overall, we recruited 976 participants (48.4% female). In the San Diego experiment ($N = 461$), we collected 111 observations in the HL treatment and 350 observations in the single-choice treatments. In Texas ($N = 515$), we had 134 participants in the HL treatment and 381 in single-choice treatments. The participants were randomly assigned to the treatments and participated in only one session. We ran all experiments at the end of an experimental session for a different study that did not relate to risk.[8] At the beginning of our experiment,

[6] Veconlab is a browser-based site for conducting economics experiments, which contains instructions and programs for many games that are commonly used in such experiments: http://veconlab.econ.virginia.edu/. It is designed and maintained by Charles Holt at the University of Virginia.

[7] There were no differences in the written instructions for the single-choice task.

[8] As a referee helpfully noted, it is possible that different experiences in the previous experiment might contribute to heterogeneity in responses to the risk elicitation tasks. Because subjects were randomly assigned to treatments, treatment effects should in principle not be biased by prior experience. However, we are unable to link data from the two experiments and so are unable to test this hypothesis.

participants received written instructions (full instructions are reported in the Electronic Supplementary Material) and were allowed to ask questions privately. Each participant then made their choices while marking Option A or B in the HL or the single-choice table in the instructions. After completing the experiment, participants were asked to come to an experimenter who rolled a ten-sided die, determining the payoffs. On the instructions sheet the experimenter also marked the gender of the participant. At the end, each participant privately received the payoff in cash and left the laboratory.

3 Results

As discussed earlier, we conducted the full HL procedure with two different sets of instructions (in the two different locations). Here we first present the results of each location separately, and then move to the joint single-choice-treatment data analysis.

We start our analysis by describing the inconsistent behavior in the HL data. As can be seen from Table 1, in the San Diego experiment, 29% of the participants violated first-order stochastic dominance (FOSD failure), by either switching more than once (Inconsistent), choosing the dominated alternative at decision 10 (Dominated) or both. Women were more likely to do so than men (36% versus 21%, respectively), and this difference is marginally significant using a two-tailed proportions test ($Z = 1.737$, $p = 0.082$).[9] Fewer participants (16%) were inconsistent in the Texas experiment, in which women again were more likely to violate first-order stochastic dominance with the same test (24% versus 11%; $Z = 1.966$, $p = 0.049$). The difference in the overall violation rate across locations is significant ($Z = -2.333$, $p = 0.020$). Pooling the data from both locations shows an overall FOSD violation rate of 22% for women and 15% for men; this difference is statistically significant ($Z = 2.784$, $p < 0.001$).

The findings in Table 1 indicate a similar direction of the gender difference with respect to inconsistency, even though different instructions and implementation lead to significantly different rates across locations. Note that the change in the instructions reduced the choices of the dominated alternative in the Texas experiment to zero. Reading the instructions aloud could be in part responsible for this; the instructions were also read aloud in the original HL experiment and in that study only one person made the dominated choice. Note that our results are consistent with those reported in Filippin and Crosetto (2016); see also footnote 5 above.

Table 2 shows the actual choices (by gender) made by all participants in San Diego and Texas, and Fig. 2 shows these in graphical form. Within each experiment, the gender differences are not statistically significant ($p = 0.482$ and $p = 0.740$ in San Diego and Texas, respectively; two-tailed Mann Whitney U test). Furthermore, none of the 20 gender comparisons are significant when we consider the ten HL rows separately and run tests of proportions for each row.[10] When pooling the data from San Diego and

[9] All statistical tests are two-tailed unless otherwise indicated.

[10] There are also no significant differences if we include only data from participants who do not violate FOSD. Detailed results are available upon request. We believe that ours is the correct comparison, because we cannot exclude single-row participants who might have been inconsistent in the full mechanism. Dropping them from only one treatment would introduce a selection effect that differs across treatments.

Table 1 Inconsistencies in 10-choice HL, by location and gender

	UC San Diego				Texas A&M			
	All [111]	Male [56]	Female [55]	*p*-value	All [134]	Male [80]	Female [54]	*p*-value
Inconsistent	0.21	0.13	0.29	0.031**	0.16	0.11	0.24	0.049**
Dominated	0.22	0.18	0.25	0.331	0.00	0.00	0.00	1.000
FOSD failure	0.29	0.21	0.36	0.082*	0.16	0.11	0.24	0.049**

The numbers of observations are in brackets. "Inconsistent" means returning to Option A after crossing over to Option B. "Dominated" stands for choosing Option A in the last row, and "FOSD failure" indicates inconsistency and/or making the dominated choice. The symbols * and ** show significance at $p = 0.10$, $p = 0.05$ level in two-tailed tests of gender differences in proportions

Texas, we find overall that the gender differences are not statistically significant ($p = 0.808$, MWU) and only one of 10 gender comparisons is significant in a proportions test.

Result 1: *We observe no gender differences in risk taking in the 10-choice HL treatment.*

Our data in both locations support the finding that females are no more risk averse than are males in the original HL task.

Considering the single-choice task, Table 3 reports the fraction of subjects choosing Option A by gender and location for the single-choice treatment. Statistical tests reported in Table 3 confirm that there is a gender difference in the lottery choices in the single-choice treatments with women typically choosing the safe Option A more often than men in both the San Diego and Texas experiments. The differences are more likely to be significant in the San Diego data. These results are illustrated graphically in Fig. 3.

We also conducted separate tests of the proportion of Option A choices, comparing choices by gender between locations for the single-choice treatments, and find that 9 out of 10 gender comparisons are not significantly different ($p > 0.1$, test of proportions) and one between-location difference is marginally significant—in Row 7 women are more risk averse in San Diego than in Texas with $p = 0.078$ (these results are available upon request). Since there is little difference across locations, we pool the data from them.

Table 4 displays the pooled results, including tests for gender differences for the pooled sample. In the pooled data, there is no difference in choices in Row 1 (perhaps due to a ceiling effect), but one must be extremely risk seeking (or confused) to choose Option B there. In Row 3, females are 18 percentage points more likely to make the safer choice and accept a reduction in expected value. In Row 5 the difference in expected value is small, and we see that females are 14 percentage points more likely to make the safer choice. In Row 7, males are much more (25 percentage points) likely to make the choice that offers a higher expected value, but has more risk. And in Row 9, females are 12 percentage points more

Table 2 Choices of Option A with 10-choice HL, by location and gender

	UC San Diego				Texas A&M				Pooled			
Row	All [111]	Male [56]	Female [55]	*p*-value	All [134]	Male [80]	Female [54]	*p*-value	All [245]	Male [136]	Female [109]	*p*-value
1	0.90	0.93	0.87	0.325	0.98	0.98	0.98	0.804	0.94	0.96	0.93	0.328
2	0.86	0.86	0.87	0.810	0.97	0.98	0.96	0.688	0.92	0.93	0.92	0.793
3	0.82	0.84	0.80	0.591	0.94	0.96	0.91	0.187	0.89	0.91	0.85	0.153
4	0.78	0.77	0.80	0.681	0.87	0.89	0.85	0.543	0.83	0.84	0.83	0.794
5	0.52	0.50	0.55	0.632	0.73	0.71	0.76	0.550	0.64	0.63	0.65	0.670
6	0.43	0.41	0.45	0.641	0.59	0.61	0.56	0.511	0.52	0.53	0.50	0.700
7	0.28	0.23	0.33	0.264	0.25	0.22	0.28	0.487	0.26	0.23	0.30	0.186
8	0.26	0.21	0.31	0.256	0.13	0.10	0.19	0.156	0.19	0.15	0.25	0.047
9	0.20	0.16	0.24	0.318	0.04	0.04	0.06	0.620	0.11	0.09	0.15	0.153
10	0.22	0.18	0.25	0.331	0.00	0.00	0.00	1.000	0.10	0.07	0.13	0.152

Percentages of participants who chose A per probability of winning the larger prize and by gender. The numbers of observations are in brackets. All *p*-values reflect M-F differences, using 2-tailed tests of proportions

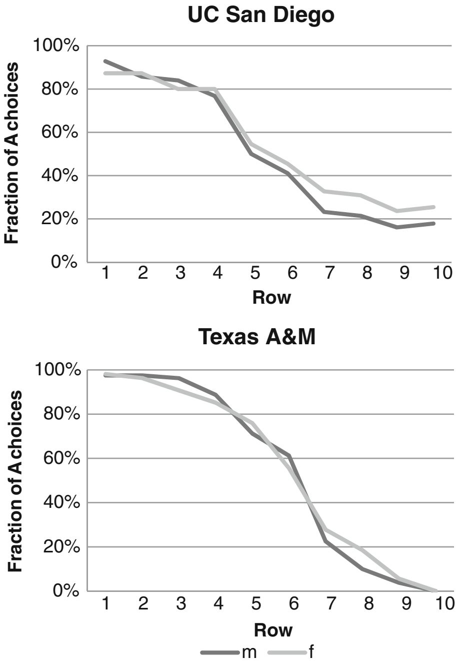

Fig. 2 Choices of Option A with 10-choice HL, by location and gender. Notes: Percentages of all participants who chose A in the full HL treatment, for each row in the HL table of decisions, by gender and location

likely to make the safer choice. Overall, Table 4 thus strongly confirms a gender difference in risk preferences, quite unlike the results with our HL treatment.

The gender differences reported in Table 4 are specified for each row separately. To measure the overall gender effect in the single-choice data, we run Probit regressions that include choices made in all single-choice treatments. Table 5 considers the likelihood of choosing Option A relative to the baseline (Row 1) in both the single-choice and the 10-row HL treatments. As can be seen, the Probit-regression results of the single-choice treatments (columns 1–3) indicate that men are significantly less likely to choose Option A than women. In the overall sample, men are 19.9 percentage

Table 3 Choices of Option A with 1-choice HL, by location and gender

	UC San Diego				Texas A&M			
Row	All	Male	Female	*p*-value	All	Male	Female	*p*-value
1	0.96 [25]	0.92 [12]	1.00 [13]	0.288	0.95 [74]	0.97 [33]	0.93 [41]	0.418
3	0.66 [100]	0.58 [50]	0.74 [50]	0.091*	0.78 [79]	0.69 [35]	0.86 [44]	0.056*
5	0.47 [100]	0.38 [50]	0.56 [50]	0.071*	0.60 [73]	0.55 [42]	0.68 [31]	0.263
7	0.34 [100]	0.18 [50]	0.50 [50]	0.001***	0.24 [78]	0.16 [37]	0.32 [41]	0.112
9	0.12 [25]	0.00 [13]	0.25 [12]	0.055*	0.12 [77]	0.09 [46]	0.16 [31]	0.319

Percentages of participants who chose A by location, labeled by the row in the original HL design, and by gender. The numbers of observations are in brackets. All *p*-values reflect the test of gender differences in proportions. *, and *** show significance at $p = 0.10$ and $p = 0.01$, respectively

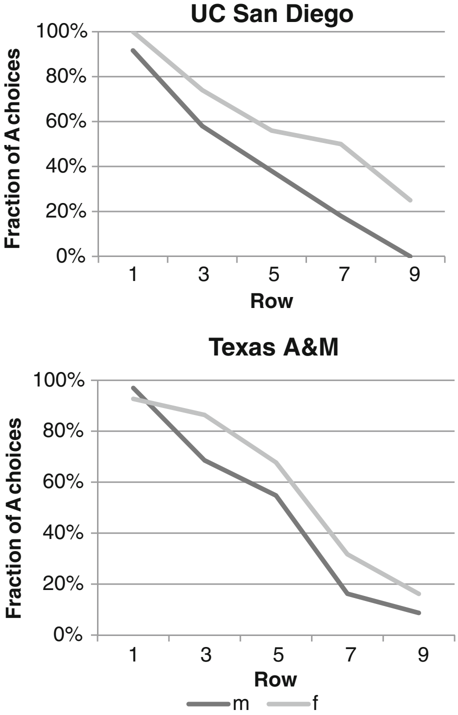

Fig. 3 Choices of Option A in the single-choice treatments, by location and gender. Notes: Percentages of all participants who chose A in the single-choice treatment for each decision, labeled by the row in the original HL table, by gender and location

points less likely to go for the safe option than women in the single-choice treatments. In the UC San Diego experiment the overall gender effect amounts to 25.2 percentage points and it is 14.6 percentage points at Texas A&M; all gender differences are highly statistically significant. By contrast, no gender differences are significant in the 10-row HL treatment (see columns 4–6).[11] The structural analyses in Appendix 1 support the results from the regressions.

Result 2: *We observe strong overall gender differences in risk taking in the single-choice treatments.*

After showing that there is no gender difference in the HL treatment, but that women are significantly more risk averse than men in the single-choice

[11] What can we say about the risk preferences implied by the single-row choices? We can calculate Coefficient of Relative Risk Aversion (CRRA) parameters for each choice. An individual who chooses Option B in Row 1 must have a CRRA coefficient less than −1.71, while someone who chooses Option B in Row 3 has a CRRA coefficient less than −0.49. An individual who chooses Option B in Row 5 must have a CRRA coefficient less than 0.14, while someone who chooses Option B in Row 7 has a CRRA coefficient less than 0.68. Finally, choosing Option B for Row 9 implies a CRRA coefficient greater than 1.37. Thus, we see from Table 4 that only 5% of the participants have a coefficient less than −1.71, 28% have a coefficient less than −0.49, 47% have a coefficient less than 0.14, 70% have a coefficient greater than 0.68, and 88% have a CRRA coefficient less than 1.37.

Table 4 Choices of Option A with 1-choice HL, by gender with pooled locations

Row	All	Male	Female	*p*-value
1	0.95 [99]	0.96 [45]	0.94 [54]	0.802
3	0.72 [179]	0.62 [85]	0.80 [94]	0.010***
5	0.53 [173]	0.46 [92]	0.60 [81]	0.051*
7	0.30 [178]	0.17 [87]	0.42 [91]	0.000***
9	0.12 [102]	0.07 [59]	0.19 [43]	0.067*

Percentages of participants who chose A per single-choice treatment and by gender. The numbers of observations are in brackets. All *p*-values represent two-tailed tests of proportions. Symbols *, and *** show significance at $p = 0.10$, and $p = 0.01$, respectively

treatments, we now compare the choices in the HL treatment with the single-choice treatments in general and also with respect to gender. Presumably, the assumption made by researchers using the HL method is that there should be no differences between the choices participants make when they are faced with the many options in the HL design and the choice made in the single-choice design. However, Table 6 shows that this assumption is rejected by the data.

Percentages of participants who chose A by decision row in 10-choice HL and single-choice treatments. The numbers of observations are in brackets. All *p*-values reflect the test of proportions. *, **, and *** show significance at $p = 0.10$, $p = 0.05$, and $p = 0.01$, respectively.

Table 5 Probit models for choices of Option A (marginal effects)

	SD-1 (1)	Texas-1 (2)	Pooled-1 (3)	SD-HL (4)	Texas-HL (5)	Pooled-HL (6)
Male	−0.252***	−0.146**	−0.199***	−0.029	−0.024	−0.026
	[0.055]	[0.060]	[0.041]	[0.063]	[0.060]	[0.043]
Row 3	−0.504***	−0.312***	−0.387***	−0.147**	−0.175**	−0.148***
	[0.136]	[0.106]	[0.081]	[0.057]	[0.075]	[0.044]
Row 5	−0.628***	−0.470***	−0.530***	−0.453***	−0.512***	−0.462***
	[0.105]	[0.085]	[0.065]	[0.059]	[0.079]	[0.046]
Row 7	−0.696***	−0.686***	−0.668***	−0.614***	−0.779***	−0.697***
	[0.086]	[0.051]	[0.047]	[0.047]	[0.043]	[0.030]
Row 9	−0.592***	−0.739***	−0.674***	−0.661***	−0.860***	−0.770***
	[0.037]	[0.041]	[0.030]	[0.041]	[0.030]	[0.024]
San Diego			−0.051			−0.072
			[0.042]			[0.044]
Observations	350	381	731	555	670	1225
Pseudo R^2	0.172	0.339	0.250	0.253	0.511	0.371

SD-1, Texas-1, and Pooled-1 represent single-choice treatments, while SD-HL, Texas-HL, and Pooled-HL represent the full 10-row HL treatment. The models are Probit specifications with the fraction of Option A choices as the dependent variable and random effects per experimental subject. Robust standard errors are clustered at the subject level. **, and *** show significance at $p = 0.05$, and $p = 0.01$, respectively

Table 6 Choices of Option A with 1-choice HL and 10-choice HL, by location

UC San Diego				Texas A&M		
Row	HL-10	1-choice	*p*-value	HL-10	1-choice	*p*-value
1	0.91 [111]	0.96 [25]	0.347	0.98 [134]	0.95 [74]	0.226
3	0.82 [111]	0.66 [100]	0.008***	0.94 [134]	0.78 [79]	0.001***
5	0.52 [111]	0.47 [100]	0.446	0.73 [134]	0.60 [73]	0.057*
7	0.28 [111]	0.34 [100]	0.340	0.25 [134]	0.24 [78]	0.965
9	0.20 [111]	0.12 [25]	0.362	0.04 [134]	0.12 [77]	0.050**

Percentages of participants who chose A by decision row in 10-choice HL and single-choice treatments. The numbers of observations are in brackets. All *p*-values reflect the test of proportions. *, **, and *** show significance at $p = 0.10$, $p = 0.05$, and $p = 0.01$, respectively

Table 6 provides a possible insight. Notice that there are more people choosing Option A in Rows 3 and 5 in HL-10 than with the single-choice—the proportions are 0.82 versus 0.66 and 0.94 versus 0.78 for Row 3 in San Diego and Texas, respectively, and 0.52 versus 0.47 and 0.73 versus 0.60 for Row 5 in San Diego and Texas, respectively. This means that people are switching later in HL-10 (where they see the full list of rows), which is consistent with the idea that they are more prone to switching to B in the middle of the table. Several related studies provide similar evidence. Andersen et al. (2006) use the HL set of decisions to investigate the effect of alternative frames that move the risk neutral switch point, and find that censoring the set of decisions moves the average elicited risk aversion level, consistent with a "middle switching" bias. Andersson et al. (2013) also vary the switch point in a list of choices, but with a different set of gambles. They too, show results consistent with some subjects switching in the middle of the list of choices.

Result 3: *We observe differences in risk taking between the HL and the single-choice treatments.*

An interesting point is that males are more affected by the change in procedure than females are. Table 7 shows the rates of Option A for males and females across HL-10 and the single-choice mechanism.

Across the HL-10 and single-choice treatments, the difference in the rates with which males choose Option A is substantially larger than the difference in the

Table 7 Choices of Option A across mechanisms, by gender

Row	Male HL-10	Male 1-choice	*p*-value	Female HL-10	Female 1-choice	*p*-value
1	0.96 [136]	0.96 [45]	0.993	0.93 [109]	0.94 [54]	0.669
3	0.91 [136]	0.62 [85]	0.001***	0.85 [109]	0.80 [94]	0.298
5	0.63 [136]	0.46 [92]	0.012**	0.65 [109]	0.60 [81]	0.512
7	0.23 [136]	0.17 [87]	0.318	0.30 [109]	0.42 [91]	0.091*
9	0.09 [136]	0.07 [59]	0.633	0.15 [109]	0.19 [43]	0.550

rates for females. Females are only slightly affected by the difference in mechanisms. It seems plausible that this effect helps drive the difference in gender effects across the single-choice and HL-10 mechanisms.

4 Conclusion

We demonstrate that reducing the complexity of an elicitation mechanism can have qualitative effects on the economic conclusions drawn. We do so by addressing a puzzle found in the literature: Why is there a consistent finding of gender differences in risk preferences with relatively simple elicitation methods such as Gneezy and Potters (1997) and Eckel and Grossman (2008a) but not with Holt and Laury (2002)? The majority of the findings in psychology and economics are in line with the simpler elicitation methods showing that, on average, women are more risk averse than men.

Is simple better? In general, more complex measures and instructions might result in a higher degree of confusion amongst the participants, and so less meaningful responses. To the extent that there are differences in results across simpler and more complex mechanisms, we believe that the results found in the simpler case reflect preferences more accurately, since the task is considerably easier to understand. While we cannot assess inconsistency across choices in our single-choice treatment, it would nevertheless seem intuitive that better-understood choices are more meaningful. There is a trade-off, of course, in that the more complex measure may also provide finer information about an individual's risk preferences. The single-row measure could not really be used as a stand-alone measure of the risk preferences of individuals, but the insights from our study can be used in the design of simpler, yet informative, alternative measures.

In closing, our results suggest that the complexity and structure of the mechanism can potentially lead researchers to draw incorrect conclusions. The issue of complexity and comprehension is also likely to be present with elicitation mechanisms in other realms of economic preferences. This is worrisome from the standpoint of the fundamentals of economic theory and we suggest that care be taken when choosing sophisticated elicitation mechanisms.

Acknowledgements We thank Chetan Dave for helping with the structural analyses, and Billur Aksoy and Kristina Bott for helping to gather the data.

Appendix 1: Structural model

An additional approach to the analysis allows us to pool all of the data into a single structural model. We assume a constant relative risk aversion utility function (CRRA), and our procedure is to obtain the structural estimates of the model as follows. We assume that subjects have a utility of money M given as $U(M|r) = \frac{M^{1-r}}{1-r}$, where r denotes the coefficient of relative risk aversion. For each binary choice between gambles, we assume that subjects evaluate the alternatives by making an expected utility calculation, weighting the utility of each outcome, $U(M_k|\, r)$, by its probability of occurrence p_k, as follows:

$$EU_i = \sum_k [p_k \times U(M_k|r)], \forall k = 1, 2$$

for each gamble i. If we further denote EU_A as the Option A gamble and EU_B as the Option B gamble, we can construct a probabilistic choice rule, where the likelihood of choosing Option A is given by:

$$\frac{EU_A^{\frac{1}{\mu}}}{EU_A^{\frac{1}{\mu}} + EU_B^{\frac{1}{\mu}}}$$

The parameter μ allows for deviations from the deterministic choice of the highest expected utility option that is specified by expected utility theory. As in previous work, μ is interpreted as a kind of noise or error: as $\mu \to \infty$ the choice becomes a random decision, and as $\mu \to 0$ subjects behave exactly as specified by expected utility theory. This parameter can also be thought of as heterogeneity in behavior not captured by the model.

The ratio above forms the basis of a logistic conditional logarithmic likelihood function, denoted as $\mathcal{L}(r, \mu|Y_i)$, which is maximized with respect to r and μ, where the vector Y_i denotes the actual subjects' choices for either Option A or Option B. In order

Table 8 Estimates of risk and noise parameters as functions of characteristics

Parameter/Variable	Estimate	St. Err.	p-value
r			
Male	−0.658	0.155	0.000
HL-10	−0.117	0.157	0.456
*HL-10*Male*	0.593	0.191	0.002
San Diego	−0.158	0.103	0.125
Constant	0.587	0.135	0.000
μ			
Male	0.122	0.089	0.170
*HL-10*Male*	−0.163	0.107	0.128
San Diego	0.168	0.059	0.004
HL-10	−0.019	0.086	0.828
Constant	0.308	0.071	0.000
Log*L*:			−1031.6
Obs.:			1956

r is a risk-aversion parameter and μ is a noise parameter. We pool all of the single-choice and corresponding HL-10 data (rows 1,3,5,7,9) for this regression. The baseline case is a female in the single-choice treatment in Texas. We cluster standard errors on the individual level to allow for heterogeneity by treatment, gender and location, each of the parameters in the vector $[r, \mu]$ is specified as a function of these factors, X_i, with associated coefficient vector β. The resulting modified likelihood function is written as $\mathcal{L}(r, \mu, \beta | Y_i, X_i)$.

The results of this estimation are shown in Table 8, which pools data from both treatments and provides a useful summary of the results discussed in the main body of the paper. The table reports a model for factors affecting the risk-aversion parameter r and the so-called "noise" parameter μ. The model includes dummy variables indicating the full HL treatment (HL-10), whether the subject is male, and if the sessions were conducted in San Diego, and an interaction between HL-10 and gender. For r, a larger positive coefficient indicates greater risk aversion, and a negative coefficient risk-seeking. For μ, larger coefficients indicate a greater divergence from Expected Utility maximization, showing degree of noise or heterogeneity not captured by the model.

The strongly negative *Male* estimated coefficient for r shows that males are considerably less risk averse than females in the single-choice treatment. Controlling for the differential effect for men, there is no significant difference in risk aversion for women across the single-choice and HL-10 mechanisms, as shown by the insignificant coefficient on HL-10, consistent with our results in Table 7. However, the large positive coefficient on the *HL-10*Male* interaction offsets the negative main effect for *Male*, indicating that the gender difference is not present with the HL-10 risk elicitation, as the sum of the *Male* and the *HL-10*Male* coefficients is an insignificant −0.065. Only the single-row treatment shows a strong gender difference. This supports our previous analysis showing a gender difference in the single-choice treatment, but not in the standard HL-10 treatment. There is also no significant difference in risk aversion across location, as indicated by the respective small-to-modest and insignificant coefficients and *San Diego*.

Turning to the noise parameter μ, the coefficient of the *Male* dummy is not significant. Furthermore, the negative coefficient of the *HL-10*Male* interaction dummy wipes out the slightly positive *Male* coefficient, so that we see no significant difference across gender with respect to noise. The slightly negative coefficient on HL-10 is far from significance. Consistent with prior analysis, we do find that the San Diego data are indeed noisier.

Note that our estimates are not strictly comparable to other structural models using HL data because we have dropped five of the "rows" in the full HL measure in order to focus on the comparison between HL and the single-row measure that we use. Nevertheless, our estimates of the HL-10 CRRA are within the range of those reported in Filippin and Crosetto (2016). The noise estimates are lower, because the dropped rows constitute a "coarser" categorization of subjects who completed the full HL. (Their study controls for differences in instructions for the task, but does not report the estimates.)

References

Andersen, S., Harrison, G., Lau, M., Rutström, E. E. (2006). Elicitation using multiple price list formats. *Experimental Economics, 9*(4), 383–405.

Andersson, O., Holm, H. J., Tyran, J. R., Wengström, E. (2013). Risk aversion relates to cognitive ability: Preferences or Noise? *Journal of the European Economic Association, 14*(5), 1129–1154.

Cason, T. N., & Plott, C. R. (2014). Misconceptions and game form recognition: Challenges to theories of revealed preference and framing. *Journal of Political Economy, 122*(6), 1235–1270.

Charness, G., & Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization, 83*, 50–58.

Charness, G., & Viceisza, A. (2016). Three risk-elicitation methods in the field: Evidence from rural Senegal. *Review of Behavioral Economics, 3*(2), 145–171.

Charness, G., Gneezy, U., Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization, 87*, 43–51.

Cox, J., Sadiraj, V., Schmidt, U. (2015). Paradoxes and mechanisms for choice under risk. *Experimental Economics, 18*(2), 215–250.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature, 47*(2), 448–474.

Dave, C., Eckel, C., Johnson, C., Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty, 41*(3), 219–243.

Eckel, C., & Grossman, P. (2008a). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization, 68*(1), 1–17.

Eckel, C., & Grossman, P. (2008b). Men, women and risk aversion: Experimental evidence. C. Plott & V. Smith (Eds.), *Handbook of Experimental Economics Results* (pp. 1061–1073). Amsterdam: North Holland.

Filippin, A., & Crosetto, P. (2016). A reconsideration of gender differences in risk attitudes. *Management Science, 62*(11), 3138–3160.

Gneezy, U., & Potters, J. (1997). An experiment on risk taking and evaluation periods. *Quarterly Journal of Economics, 102*, 631–645.

Halevy, Y. (2007). Ellsberg revisited: An experimental study. *Econometrica, 75*(2), 503–536.

Healy, P. J., & Brown, A. (2016). Separated decisions. Working papers 16-02. Columbus: Ohio State University, Department of Economics.

Hey, J., Morone, A., Schmidt, U. (2009). Noise and bias in eliciting preferences. *Journal of Risk and Uncertainty, 39*(3), 213–235.

Holt, C., & Laury, S. (2002). Risk aversion and incentive effects. *American Economic Review, 92*(5), 1644–1655.

Holt, C., & Laury, S. (2014). Assessment and estimation of risk preferences. In M. Machina & W. K. Viscusi (Eds.), *Handbook of the Economics of Risk and Uncertainty* (pp. 135–201). Oxford: North Holland.

Miller, L., Meyer, D. E., Lanzetta, J. T. (1969). Choice among equal expected value alternatives: Sequential effects of winning probability level on risk preferences. *Journal of Experimental Psychology, 79*(3), 419–423.